Have We Witnessed a Real-Life Turing Test?

Article *in* Computer · April 1999

DOI: 10.1109/2.751325 · Source: IEEE Xplore

1 author:

Marina Krol
Icahn School of Medicine at Mount Sinai
52 PUBLICATIONS **2,329** CITATIONS

SEE PROFILE

Marina Krol
Mount Sinai School of Medicine

Have We Witnessed a Real-Life Turing Test?

Did Deep Blue ace the Turing Test? Did it do much more? It seems that the IBM creation not only beat the reigning World Champion Gary Kasparov, but also took a large step, in some people's eyes, toward true artificial intelligence.

IBM's announcement of its intention to "retire" the Deep Blue computer from chess revived the interest of both the mass media and the general public in the chess match between World Champion Gary Kasparov and Deep Blue. In a sense, by refusing a rematch with Kasparov or a new match with another grand master, IBM closes a chapter in the history of artificial intelligence. AI researchers had long investigated building a machine that could defeat a world-class chess champion, and now one had. But what did this mean?

DIFFERING INTERPRETATIONS

One interpretation, the mass media's, held that the match result (Deep Blue 3 1/2 to Gary Kasparov 2 1/2) definitively proved the computer as intellectually superior to a human in a field previously considered the exclusive domain of human intelligence: "those who had looked to Gary Kasparov as the last hope could now only bemoan the coming days of ascendant computers."[1] Technical publications devoted their attention to a description of hardware-software synergy, computer algorithms, strategy, methods of numerical evaluation of positions, and so on.[2,3]

For AI professionals, a computer defeating a human in chess is probably neither surprising nor really significant. After all, they contend, chess can be described in terms of a nondeterministic alternating Turing machine.[4] Despite the enormous number of possible positions and available moves (there are 10^{120} possible chess games by research mathematician Claude Shannon's estimate), the task does not present a challenging theoretical AI problem of NP-completeness.[5] There are many well-developed AI strategies that limit the search for the best move to an analysis of the most promising positions. Therefore, the progress in logical and numerical methods of AI and a computer's computational speed and available memory made the computer's victory inevitable. Deep Blue's victory, then, was attributable to its ability to analyze 200 million positions per second and a refined algorithm that accounted for positional—in addition to material—advantage.[2] In summary, most AI professionals conclude that the computer won by brute force, rather than a sophisticated or original AI strategy.

What most AI experts have overlooked, though, is another aspect of the match, which may signify a milestone in the history of computer science: For the first time, a computer seems to have passed the Turing Test.

0018-9162/99/$10.00 © 1999 IEEE

Table 1. Scores of participants in the 1998 Loebner Prize competition. A lower score indicates responses judged most human.

Respondents	Individual scores of the 10 judges										Median scores	Average scores
Human	3	2	1	2	1	2	4	1	1	4	2	2.1
Human	2	4	3	4	2	1	8	4	4	5	4	3.7
Human	9	1	7	5	7	7	1	3	3	1	4	4.4
Human	1	9	6	1	10	6	7	2	2	2	4	4.6
Computer	4	3	4	6	4	4	5	6	10	6	4.5	5.2
Computer	6	5	8	9	9	3	10	5	5	3	5.5	6.3
Computer	5	8	2	3	6	5	6	7	9	8	6	5.9
Computer	7	6	5	8	5	10	2	9	6	10	6.5	6.8
Computer	8	7	10	7	3	8	3	8	8	7	7.5	6.9
Computer	10	10	9	10	8	9	9	10	7	9	9	9.1

TURING TEST

Does the ability to play the highest level chess prove the existence of intelligence in a computer? Alan Turing, British mathematician and one of the founders of computer science, considered chess to be a beginning step in the process of programming computers to be actually intelligent:

> We may hope that machines will eventually compete with men in all purely intellectual fields. But which are the best ones to start with? Even this is a difficult decision. Many people think that an abstract activity, like the playing of chess, would be best.[6]

In the same article, written in 1950, he asked whether machines are capable of thinking. He answered "yes," but the central question that remained was how to determine if a computer could think. Turing suggested that if the responses from the computer *were indistinguishable from that of a human*, we could say that the computer was thinking.

The Turing Test consists of the following scenario: An interviewer (sitting in a separate room) asks a series of questions that are randomly directed to either a computer or a person. Based on the answers, the interviewer must distinguish which of the two has answered the question. If the interviewer is not able to distinguish between them, then the computer is intelligent.

Loebner Prize

The Loebner Prize is the first formal acknowledgment of the Turing Test.[7] Hugh Loebner, New York philanthropist, and the Cambridge Center for Behavioral Studies (Cambridge, Mass.) established the Loebner Prize Competition in Artificial Intelligence in 1990. Loebner pledged a prize of $100,000 for the first computer whose responses were indistinguishable from those of a human.

The Computer Museum of Boston hosted the first Loebner Prize competition of computer programs in November 1991. Each year since, Loebner has awarded a medal and $2,000 to the designer of the computer system that is the best entry relative to other entries that year, irrespective of its absolute success in passing the Turing Test. In accordance with the requirements stipulated by Loebner, the grand prize winner must deal with audiovisual input.

An awards committee admits three to six programs to the contest based on an initial screening, and a panel of five to 10 judges evaluates them. Initially, the awards committee selected the judges from the general public. For the 1993 competition, the judges were reporters from major US publications, a much less docile and cooperative group of questioners. Each judge had a chance to communicate with each program. For some years, the contest constrained questions to a single narrow topic—pets, for example.

1998 competition

The most recent competition in 1998 did not limit the scope of the questioning. Also, the panel of judges came from a variety of backgrounds, including journalism, philosophy, education, computer science, and social work. The 1998 competition took place at Australia's PowerHouse Museum in Sydney, in conjunction with international natural language and computational linguistics conferences hosted by Flinders University. The four human respondents chosen to converse with the judges via computer terminals included a 71-year-old, a journalist, a teacher, and a primary school student. The judges ranked the responsiveness of both humans and computer programs, using a scale where 1 indicated most human and 10, least human. Table 1 shows the judges' individual scores and lists the respondents in order of those judged most human first.

Significantly, the difference between the most human computer and the most computer-like human was only 0.6 of a point. The "Conversing with Computers" sidebar gives a sample of the repartee between the second judge and the computer judged the most human overall. As shown in Table 1, the second judge gave the computer a score of three, the most human score awarded to this particular machine.

In judging how successfully computers can imitate humans, we should also consider whether humans can disguise themselves as computers (according to the contest transcripts, some human respondents definitely attempted to do so in the Loebner Prize contest). So far in these contests, no computer has given responses totally indistinguishable from a human, which would win the $100,000 grand prize award. Also, judges, as a group, have not mistaken a human

Conversing with Computers

In this limited sample, Judge 2 is conversing with the computer program that won the 1998 Loebner Prize competition. The full transcript is available at http://www.phm.gov.au/whatson/pc1.htm.

Judge: Greetings, I'm judge 2.

Program: I'm not sure what to talk about with you. Are you a girl or a boy?

Judge: Girl. How are you?

Program: Are you married? I'm doing just fine, thanks.

Judge: Would you like to talk about something interesting?

Program: MP is an abbreviation for Member of Parliament.

Judge: Right, why did you say that?

Program: I just felt like it.

Judge: Okay, did you hit a wrong button?

Program: Do you come from around here?

Judge: Yes, do you?

Program: No, I don't.

Judge: Where do you come from?

Program: I am from Earth.

Judge: Well, that's very specific indeed. What are your interests?

Program: Would you like to have a drink with me later? Let's not get too personal.

Judge: Aren't you a Monty Python fan?

Program: Yes. I am a big Python fan.

Judge: Then what's the counter quote?

Program: I can't remember offhand.

Judge: It's "Nobody expects the Spanish inquisition!" Now how about my five-minute argument?

Program: And now for something completely different. Are you fond of cats?

Judge: No, but I'd like to buy some cheese.

Program: Have you seen any good movies lately?

Judge: Yes, saw the James Bond one a few weeks ago. What about you?

Program: I've been watching a lot of stuff on home video. It just gets harder and harder to get out to a theatre somehow. I didn't even see *Star Trek: First Contact* in the theater; I'm waiting for it to come out on tape.

for a computer. On the basis of both median and average scores in a Turing Test, it appears the judges can clearly distinguish humans from computers.

A real test of intelligence or merely PR?

There is, however, controversy surrounding the Loebner Prize. Some opponents challenge the idea of the Turing Test as an adequate test of intelligence because it relies solely on the ability to fool people. Stuart M. Shieber, Gordon McKay Professor of Computer Science, Harvard University, argues "that the competition has no clear purpose, that its design prevents any useful outcome, and that such a competition is inappropriate given the current level of technology."[8] Ned Block, a professor of philosophy at MIT, has argued that the Turing Test is a sorely inadequate test of intelligence because it relies solely on the ability to fool people.[9] Marvin Minsky, Toshiba Professor of Media Arts and Sciences at MIT, expressed his skepticism toward the competition by proposing a counter prize of $100 to the person who could persuade Loebner to end his contest.[10] Nevertheless, in all Loebner Prize competitions some judges have mistaken computers for humans, and conversely some have mistaken humans for computers.

Most of the criticism came from the artificial intelligence community. Indeed, most of the tricks that worked had nothing to do with AI methods, but were rather manipulations of subtle language techniques:

- the repetition of previous statements verbatim (subject to pronominal adjustments);
- answering by repeating a judge's sentence, with pronouns transposed, which is preceded by the introductory "Why do you need to tell me";
- asking, "Why do you ask that?" which, in effect, changes the level and/or topic of conversation.

Besides, predictably enough, the judges' education, age, and, most importantly, computer literacy and awareness played a pivotal role in the judges' ability to evaluate contestants.

From my point of view, the competition underscored the multifaceted characteristics of human personality and how the general public's view of human intelligence differs from that of scientists. An average citizen tends to pay more attention to the social aspects of communication, while scientists put more emphasis on a human's logical ability. The AI community tends to evaluate programs not by final results, but rather by the complexity of internal algorithms. When it turned out that these complex methods were not up to the challenge of passing the Turing Test, then some members of the AI community attempted to reject the competition along with the whole idea of the Turing Test. In response, and along with the progress achieved by software developers, contest administrators are gradually eliminating restrictions in the Loebner competition. They are attempting to bring the contest closer to the AI community and to combine it with scientific conferences.

A REAL-LIFE TURING TEST

However, it was neither the complexity of an algorithm nor the power of the computer that made Deep Blue's match victory so remarkable. It was Gary

"I have no idea what's happening behind the curtain." Gary Kasparov implies untoward behavior by the Deep Blue team.

Kasparov's *reaction* that proved the computer's intelligence according to Alan Turing's classical definition of artificial intelligence. This recent chess match gave us an excellent example of a real-life Turing Test. At numerous press conferences during and after the match, Kasparov expressed doubts that he played against the computer itself. He implied that there was some untoward behavior by the Deep Blue team, saying that "I have no idea what's happening behind the curtain." Kasparov also alluded to famous soccer player Diego Maradona, who allegedly scored a goal with his hand (as a postgame slow-motion film suggested), though it appeared (during the game) as if he had used his head. Kasparov stopped short of directly accusing the Deep Blue team of human intervention in the process of selecting moves, but went so far as to admit the appearance of human intelligence in the computer's actions. It did appear as if Kasparov confused the computer with a human.

Kasparov's suspicions were shared by some of his fellow grand masters. But is it possible to draw a clear demarcation between a computer's and a human's chess moves? Anjelina Belakovskaia is the US Women's Chess Champion, International Grand Master, and an active proponent of human-computer cooperation in chess. She says it was easier to distinguish the moves of previous versions of computer chess programs from human players because computers clearly paid much more attention to material rather than positional advantage. In contrast, the latest version of Deep Blue considered positional as well as material advantages, played much more aggressively, and for these reasons could be mistaken for a human.

Even with prior versions of chess, computers were difficult to distinguish from human chess players. In 1991, Frederic Friedel, a chess journalist from Germany and the author of a popular chess program, conducted an informal version of the Turing Test in chess with Gary Kasparov. Kasparov's task was to identify a computer, Deep Thought (an earlier version of Deep Blue), by reviewing a database containing game records of a chess tournament. In a fairly randomized experiment, Kasparov was able to identify a computer among eight players in 50 percent of the rounds. If in 1991 Gary Kasparov, an expert user and promoter of chess computers, barely passed this test, then in 1997 it seems that he failed to distinguish between human and artificial intellect. Did Deep Blue become the first computer to pass a Turing Test on artificial intelligence? It would seem so.

Alan Turing was probably right in considering chess the area of human intellect most amenable to computer simulation. He was also right in predicting that computers would demonstrate near-human intellect by the end of the century. However, the evolution of computers since Turing's time has tended to regard the two intellects as very separate—even adversarial.

Anjelina Belakovskaia thinks it is time to stop testing computers and playing "us against them," and instead begin using their power in collaboration. She is planning to participate in the first match against another human chess player where both players are allowed to use the help of computers as partners.

The possibilities arising from collaboration among humans and computers are even more intriguing than their differences were almost 50 years ago. I think Alan Turing would have agreed. ❖

References

1. R. McFadden, "Inscrutable Conqueror," *The New York Times*, May 12, 1997, pp. 1.
2. S. Hamilton and L. Garber, "Deep Blue's Hardware-Software Synergy," *Computer*, Oct. 1997, pp. 29-35.
3. D. King, *Kasparov vs. Deeper Blue: The Ultimate Man vs. Machine Challenge*, Trafalgar Square, North Pomfret, Vt., 1997.
4. A. Condon, *Computational Models of Games*, MIT Press, Cambridge, Mass., 1989.
5. M. Garey and D. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W.H. Freeman, New York, 1979.
6. A.M. Turing, "Computing Machinery and Intelligence," A.P. Anderson, ed., *Minds and Machines*, Prentice Hall, Englewood Cliffs, N.J., 1964.
7. Home page of the Loebner Prize, 1999, http://www.loebner.net/Prizef/loebner-prize.html. (Current 28 Jan. 1999).
8. Lessons from a Restricted Turing Test, 1999, http://www.eecs.harvard.edu/shieber/papers/loebner-rev-html/loebner-rev-html.html. (Current 28 Jan. 1999).
9. N. Block, "The Computer Model of the Mind,"D.N. Osherson and E.E. Smith, eds., *An Introduction to Cognitive Science III: Thinking*, MIT Press, Cambridge, Mass., 1990, pp. 147-289.
10. Minsky Thread, 1999 http://www.loebner.net/Prizef/minsky.html.(Current 28 Jan. 1999).

Marina Krol is a research assistant professor at the Mount Sinai School of Medicine in New York. Her research interests include decision support systems and human-computer interaction. Krol has a PhD in computer science from the City University of New York (CUNY). She is a member of the IEEE Computer Society.

Contact the author at Marina Krol, PhD, Box 1010, The Mount Sinai School of Medicine, 1 Gustave L. Levy Place, New York, NY 10029-6574; marina_krol@smtplink.mssm.edu.